MTH302 Short Notes for Final Term (Chapter 23-45)

STATISTICAL DATA:

Information is collected by government departments, market researchers, opinion pollsters and others.

BASIS FOR CLASSIFICATION:

1. Qualitative: Attributes: sex, religion
2. Quantitative Characteristics: Heights, weights, incomes etc.
3. Geographical: Regions: Provinces, divisions etc.
4. Chronological or Temporal
5. By time of occurrence: Time series

TYPES OF CLASSIFICATION:

There are different types of classifications.

- One-way
  One characteristic: Population
- Two-way
  Two characteristics at a time
- Three-way
  Three characteristics at a time

METHODS OF PRESENTATION:

Different methods of representation are:

- Text
  ”The majority of population of Punjab is located in rural areas.”
- Semi-tabular
  Data in rows
- Tabular
  Tables with rows and columns
- Graphic
  Charts and graphs

TYPES OF GRAPHS:

- Column Graphs
- Line Graphs
- Circle Graphs (Sector Graphs)
- Conversion Graphs
- Travel Graphs
- Statistical Graphs
- Frequency Tables

- Histograms
- Frequency distributions
- Cumulative Distributions

LINE GRAPHS:

Line graphs are the most commonly used graphs. Here the data of one variable (say Height) is plotted against data of the other variable (say Age).

MEAN:

The most common average is the mean. The mean is used for things like marks and scores (e.g. sport), and is found by adding all the scores and dividing by the number of scores.

Example:

Marks
58 69 73 67 76 88 91 and 74 (8 marks).
Sum = 596
Mean = 596/8 = 74.5

Please note that the mean is affected by extreme values.

MEDIAN:

Another typical value is the median. The median is the middle value when the data are arranged in order.The median is easier to find than the mean, and unlike the mean it is not affected by values that are unusually high or low

Example:

Data
3 6 11 14 19 19 21 24 31 (9 values)

The median is the middle score, or the mean of the two middle scores, when the scores are placed in order. In the above data there are 9 values. The middle value is 19.

When there is no middle value, the median is obtained by taking the average of the two middle values.

MODE:

The most common score in a set of scores is called the mode.
There may be more than one mode, or no mode at all
2 2 1 2 0 3 2 1 1 4 1 1 1 2 2 0 3 2 1

The mode, or most common value, is 1.

ORGANISING DATA:

There are many different ways of organizing data.

Organizing Numerical Data

Numerical data can be organized in any of the following forms:

- The Ordered Array and Stem-leaf Display
- Tabulating and Graphing Numerical Data
- Frequency Distributions: Tables, Histograms, Polygons
- Cumulative Distributions: Tables, the Ogive

Tabulating and Graphing Univariate Categorical Data:

There are different ways of organizing univariate categorical data:

- The Summary Table
- Bar and Pie Charts, the Pareto Diagram

Tabulating and Graphing Bivariate Categorical Data:

Bivariate categorical data can be organized as:

- Contingency Tables
- Side by Side Bar charts

GRAPHICAL EXCELLENCE AND COMMON ERRORS IN PRESENTING DATA:

It is important that data is organized in a professional manner and graphical excellence is achieved in its presentation.

Tabulating Numerical Data:

The process of developing frequency distributions is described below.

Step 1: Sort Raw Data in Ascending Order

Data: 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Step 2: Find Range
Range: 58 - 12 = 46

Step 3: Select Number of Classes
Select the number of classes. (The classes are usually selected between 5 and 15)

Step 4: Compute Class Interval (width)
Step 5: Determine Class Boundaries (limits)

Start with 10 as the first limit.
Then add 10 to each limit: 10, 20(=10+10), 30(=20+10), 40(=30+10), 50(=40+10)

Step 6: Compute Class Midpoints

First midpoint is 10+20/2=15.

Midpoints: 15((10+20)/2), 25((20+30)/2), 35((30+40)/2), 45((40+50)/2), 55((50+60)/2)

Step 7: Count Observations & Assign to Classes

First class: Lower limit is 10. Higher limit is 20. We read it as "10 but under 20". In reality a value greater than 19.5 will be treated as above 20. Frequency: Looking through the data shows that there are three values between 10 and 20. Hence frequency is 3. Similarly, frequency in other intervals can be found as follows:

20 - 30 : 6

30 - 40 : 5

40 - 50 : 4

50 - 60 : 2

Total : 20

Relative frequency:

Relative frequency of a class = $\frac{\text{Frequency of the class interval}}{\text{Total Frequency}}$

There are 3 observations in class interval 10 – 20. The relative frequency is 3/20 = 0.15.

Percentage Frequency:

If we multiply 0.15 by 100, then the % Relative Frequency 15% is obtained.

Cumulative Frequency:

If we add frequency of the second interval to the frequency of the second interval, then the cumulative frequency for the second interval is obtained. The cumulative frequency of the last interval is 100% as all observations have been added.

GRAPHING NUMERICAL DATA: THE HISTOGRAM:

When frequency is plotted in the form of bars or columns for each class interval a Histogram is obtained. (Average, Averagea)

MEASURES OF CENTRAL TENDENCY:

Measures of central tendency can be summarized as under:

- Arithmetic Mean
- Arithmetic Mean for Grouped Data
- Weighted Mean
- Median
- Median for Grouped Data
- Median for Discrete Data
- Graphic Location of Median
- Quintiles (Quartiles, Deciles, Percentiles)
- Quintiles from Grouped Data
- Quintiles from Discrete Data
- Graphic Location of Quintiles
- Mode
- Mode from Grouped Data
- Mode from Discrete Data
- Empirical Relation Between mean, Median and Mode

AVERAGE:

Returns the average (arithmetic mean) of the arguments. AVERAGE (number1,number2,...) Number1, number2, ... are 1 to 30 numeric arguments for which you want the average.

AVERAGEA:

Calculates the average (arithmetic mean) of the values in the list of arguments. In addition to numbers, text and logical values such as TRUE and FALSE are included in the calculation.
AVERAGEA(value1,value2,...)

MEDIAN:

The median is the number in the middle of a set of numbers; that is, half the numbers have values that are greater than the median, and half have values that are less.
MEDIAN(number1,number2,...)

MODE:

Returns the most frequently occurring, or repetitive, value in an array or range of data. Like MEDIAN, MODE is a location measure.
MODE(number1,number2,...)

COUNT FUNCTION:

Counts the number of cells that contain numbers and also numbers within the list of arguments. Use COUNT to get the number of entries in a number field that's in a range or array of numbers.
COUNT(value1,value2,...)

FREQUENCY:

Calculates how often values occur within a range of values, and then returns a vertical array of numbers. For example, use FREQUENCY to count the number of test scores that fall within ranges of scores. Because FREQUENCY returns an array, it must be entered as an array formula.
FREQUENCY(data_array,bins_array)

ARITHMETIC MEAN GROUPED DATA:

Below is an example of calculating arithmetic mean of grouped data. Here the marks and frequency are given. The class marks are the mid points calculated as average of lower and higher limits. For example, the average of 20 and 24 is 22. The frequency f is multiplied by the class mark to obtain the total number. In first row the value of fx is 1 x 22 = 22. The sum of all fx is 1950. The total number of observations is 50. Hence the arithmetic mean is 1950/50 = 39.

Marks	Frequency	Class Marks	fX
20-24	1	22	22
24-29	4	27	108
30-34	8	32	256
35-39	11	37	407
40-44	15	42	630
45-49	9	47	423
50-54	2	52	104
TOTAL	50		1950

n= 50; Sum(fX) =1950; Mean =1950/50= 39 Marks

CUMULATIVE FREQUENCY:

Relative frequency can be converted into cumulative frequency by adding the current frequency to the previous total.
RF + CF = Cumulative Frequency

Percent Cumulative frequency is calculated by dividing the cumulative frequency by the total number of observations and multiplying by 100. For the first interval the % cumulative frequency is 3/20*100 = 15%

CUMULATIVE % POLYGON-OGIVE:

From the % cumulative frequency polygon that starts from the first limit (not mid point as in the case of relative frequency polygons) can be drawn. Such a polygon is called Ogive. The maximum value in an Ogive is always 100%. Ogives are determining cumulative frequencies at different values (not limits).

TABULATING AND GRAPHING UNIVARIATE DATA:

Univariate data (one variable) can be tabulated in Summary form or in graphical form. Three types of charts, namely, Bar Charts, Pie Charts or Pareto Diagrams can be prepared.

GEOMETRIC MEAN:

Geometric mean is defined as the root of product of individual values. Typical syntax is as under:
G=(x1.x2.x3 ... xn)^1/n

Example:
Find GM of 130, 140, 160

GM = (130*140*160)^1/3
= 142.8

HARMONIC MEAN:

Harmonic mean is defines as under:
HM=n / (1/x1+1/x2+.... 1/xn)
= n / Sum(1/xi)

Example:
Find HM of 10, 8, 6
HM = 3/(1/10+1/8+ 1/6)
= 7.66

QUARTILES:

Quartiles divide data into 4 equal parts

1st Quartile Q1=1 (n+1)/4
2nd Quartile Q2= 2(n+1)/4
3rd Quartile Q3= 3(n+1)/4

Grouped data

Qi= ith Quartile = l + h / f [Sum f / 4*i – cf)
l = lower boundary
h = width of CI
cf = cumulative frequency

DECILES:

Deciles divide data into 10 equal parts

1st Decile D1= (n+1)/10

2nd Decile D2= 2(n+1)/10 9th

Deciled D9= 9(n+1)/10 Grouped data

Qi = ith Decile (i=1,2,.9) = l + h/f[Sum f/10*i – cf)
l = lower boundary
h = width of CI
cf = cumulative frequency

PERCENTILES:

Percentiles divide data into 100 equal parts

1st Percentile P1=(n+1)/100
2nd Decile D2= 2(n+1)/100
99th Deciled D9= 99(n+1)/100

Grouped data

Qi = ith Decile(i=1,2,.9) = l + h/f[Sum f/100*i – cf)
l = lower boundary
h = width of CI
cf = cumulative frequency

EMPIRICAL RELATIONSHIPS:

Symmetrical Distribution
mean = median = mode

Positively Skewed Distribution
(Tilted to left)
mean > median > mode

Negatively Skewed Distribution
mode > median > mean
(Tilted to right)

Moderately Skewed and Unimodal Distribution

Mean – Mode = 3(Mean – Median)

Example:

mode = 15, mean = 18, median = ?

Median = 1/3[mode + 2 mean]

= 1/3[15 + 2(18)]

= [15+36]/3 = 51/3 = 17

mean = (mean – median / 3) + mode

mode = (mean – median /3) – mean

median= mode + 2mean/3

Winsorized MEAN:

Replace each observation below first quartile with value of first quartile Replace each observation above the third quartile with value of 3rd quartile

TRIMMED AND WINSORIZED MEAN:

Example:

Find trimmed and winsorized mean.

9.1, 9.1, 9.2, 9.3, 9.2, 9.2

Array the data

9.1, 9.1, 9.2, 9.2, 9.2, 9.2, 9.3,

Q1 = (6+1)/4=1.75 (2nd value) = 9.2

Q3 = 3(6+1)/4= 5.25 (6th value) = 9.3

TM= (9.2+9.2+9.2+9.2+9.3)/5 = 9.22

WM = (9.2+ 9.2+9.2+9.2+9.2+9.3+9.3)/7

DISPERSION OF DATA:

The degree to which numerical data tend to spread about an average is called the dispersion of data

TYPES OF MEASURES OF DISPERSION:

Absolute measures

Relative measures (coefficients)

DISPERSION OF DATA:

Types of Absolute Measures:

- Range
- Quartile Deviation
- Mean Deviation
- Standard Deviation or Variance

Types of Relative Measures

- Coefficient of Range

- Coefficient of Quartile Deviation
- Coefficient of Mean Deviation
- Coefficient of Variation MEASURES OF CENTRAL TENDENCY, VARIATION AND SHAPE FOR A SAMPLE

MEANS:

The most common measure of central tendency is the mean.

THE MEAN:

It is the sum of all values divided by the number. In the case of mean of a sample, the number n is the total sample size.

EXTREME VALUES:

An important point to remember is that arithmetic mean is affected by extreme values.

THE MEDIAN:

The Median is derived after ordering the array in ascending order. If the number of observations is odd, it is the middle value otherwise it is the the average of the the two middle values. It is not affected by extreme values.

THE MODE:

The mode is the value that occurs most frequently.

MIDRANGE:

Midrange is the average of slimmest and largest value. In other words it is half of a range. Midrange is affected by extreme values as it is based on smallest and largest values

QUARTILES:

Quartiles are not exclusively measures of central tendency. However, they are useful for dividing the data in 4 equal parts.

QUARTILE DEVIATION:

Quartile Deviation is the average of 1st and 3rd Quartile.

$$Q.D = (Q3 - Q1)/2$$

Example

Find Q.D

14, 10, 17, 5, 9, 20, 8, 24, 22, 13

Q1 = (n+1)/4th value = (10+1)/4 = 2.75th

= 8 + 0.75(9 - 8)= 8 + 0.75 x 1= 8.75

Q3= 3(2.75) = 8.25th value

= 8th value + 0.25(9th value –8th value)

= 20 + 0.25 (22 –20) = 20.50

Q.D =(20.50 –8.75)/2 = 5.875

BOX AND WHISKER PLOTS:

Box and Whisker plots show the 5 number summary:

- Smallest value
- 1st Quartile (Q1)
- Median(Q2)
- 3rd Quartile (Q3)
- Largest value

MEASURES OF VARIATION:

In measures of variation, there are the sample and population standards deviation and variance the most important measures. The coefficient of variation is the ratio of standard deviation to the mean in %.

INTERQUARTILE RANGE:

Interquartile range is the difference between the ist and 3rd quartile.

VARIANCE:

Variance is the one of the most important measures of dispersion. Variance gives the average square of deviations from the mean. In the case of the population, the sum of square of deviations is divided by N the number of values in the population. In the case of variance for the sample the number of observations less 1 is used.

STANDARD DEVIATION:

Standard deviation is the most important and widely used measure of dispersion. The square root of square of deviations divided by the number of values for the population and number of observations less 1 gives the standard deviation.

COEFFICIENT OF VARIATION:

Coefficient of variation (CV) shows the dispersion of the standard deviation about the mean. In the slide you see two stocks A and B with CV=10% and 5% respectively.

MEAN DEVIATION:

Other useful measures are Deviation about the Mean and median. The formulas

for normal or grouped data are as follows:

Mean Deviation About Mean – Normal data

MD (mean) = Sum (xi- mean)/n

For Grouped data – Grouped data
MD (mean) = Sum fi (xi - mean)/Sum fi

Mean Deviation About Median – Normal data
MD (median) = Sum (xi-median)/n

Mean Deviation About Median – Grouped data

MD (median) = Sum fi (xi-median)/Sum fi

CORRELATION:

The main function of regression analysis is determining the simple linear regression equation.

SCATTER DIAGRAM:

The first step in regression analysis is to plot the values of the dependent and independent variable in the form of a scatter diagram

Types of Regression Models:

There are two types of linear regression models. These are positive and negative linear relationships. In the positive relationship, the value of the dependent variable increases as the value of the independent variable increases. In the case of negative linear relationship, the value of the dependent variable decreases with increase in the value of independent variable.

INTERCEPT:

Use the INTERCEPT function when you want to determine the value of the dependent variable when the independent variable is 0 (zero). For example, you can use the INTERCEPT function to predict a metal's electrical resistance at 0°C when your data points were taken at room temperature and higher.

Types of Regression Models:

There are different types of regression models. The simplest is the Simple Linear Regression Model or a relationship between variables that can be represented by a straight line equation.

REGRESSION EQUATION

The formula for the regression equation is as under:

Equation of Least Squares Regression line

y – ym = (r.s(y)/s(x)).(x-xm)

Example

Based on analysis of data the following values have been worked out:

xm = 4;

ym = 80;

s(x) = 2^1/2;

s(y) = 200^1/2;

r = 0.8

Find the regression equation Y = a + b.X

Using the formula given above:
y – ym = (r.s(y)/s(x)).(x-xm)
y – 80 = 0.8 . 200^1/2 . (x – 4)
2^1/2
y – 80 = 8(x – 4)
y = 8x – 32 + 80
y = 8x + 48

SAMPLING DISTRIBUTION:

It is possible to construct a sampling distribution for r similar to those for sampling distributions for means and percentages.

SEASONABLE VARIATIONS:

Seasonal Variations are regarded as constant amount added to or subtracted from the trends. This is a reasonable assumption as seasonal peaks and troughs are roughly of constant size.

PERMUTATIONS

An arrangement of all or some of a set of objects in a definite order is called permutation.

BINOMDIST

Returns the individual term binomial distribution probability. Use BINOMDIST in problems with a fixed number of tests or trials.

BINOMDIST(number_s,trials,probability_s,cumulative)
Number_s is the number of successes in trials.
Trials is the number of independent trials.
Probability_s is the probability of success on each trial.
Cumulative is a logical value that determines the form of the function.

NEGBINOMDIST:

Returns the negative binomial distribution. NEGBINOMDIST returns the probability that there will be number_f failures before the number_s-th success, when the constant probability of a success is probability_s.

CRITBINOM:

Returns the smallest value for which the cumulative binomial distribution is greater than or equal to a criterion value.

POISSON DISTRIBUTION:

The POISSION Distribution has the following characteristics:

- Either or situation
- No data on trials
- No data on successes
- Average or mean value of successes or failures

STANDARDISATION:

Process of calculating z from x is called Standardisation. z indicates how many standard deviations the point is from the mean

SAMPLING DISTRIBUTION:

The sampling distribution of percentages is the distribution obtained by taking all possible samples of fixed size n from a population

Mean of the Sampling Distribution:

Using the above data:

Mean = 100% x 0.2 + 67% x 0.6 + 33% x 0.2 = 67%

DISTRIBUTION OF SAMPLE MEANS:

The standard deviation of the Sampling Distribution of means is called Standard Error of the Mean STEM.

STEM = s.d/(n)^1/2

s.d denotes standard deviation of the population.
n is the size of the sample.
